SCIENCE CHINA
Life Sciences

• RESEARCH PAPER •

July 2015 Vol.58 No.7: 682–686
doi: 10.1007/s11427-015-4857-9

Non-coding regions of the Ebola virus genome contain indispensable phylogenetic and evolutionary information

JIANG XinQuan[1], ZHANG ZhenJie[2], ZHUANG DongMing[2], CARR Michael J.[3], ZHANG RuiLing[2], LV Qiang[2] & SHI WeiFeng[2*]

[1]*School of Public Health, Taishan Medical College, Tai'an 271000, China;*
[2]*Institute of Pathogen Biology, Taishan Medical College, Tai'an 271000, China;*
[3]*National Virus Reference Laboratory, University College Dublin, Dublin 4, Ireland*

Received March 10, 2015; accepted April 2, 2015; published online May 6, 2015

We compared the numbers of nucleotide substitutions occurring in the non-coding regions and coding regions of Ebola virus genomes and found that non-coding regions contain indispensable phylogenetic and evolutionary information. The omission of genetic data from non-coding regions can lead to unreliable phylogenies and inaccurate estimates of evolutionary parameters.

EBOV, Ebola, non-coding region, phylogenetic analysis, evolutionary information

Citation: Jiang XQ, Zhang ZJ, Zhuang DM, Carr MJ, Zhang RL, Lv Q, Shi WF. Non-coding regions of the Ebola virus genome contain indispensable phylogenetic and evolutionary information. Sci China Life Sci, 2015, 58: 682−686, doi: 10.1007/s11427-015-4857-9

In March 2014, the World Health Organization (WHO) was notified of an Ebola virus (EBOV) outbreak in Guinea [1]. To March 15, 2015, there are 24,701 confirmed, suspected or probable cases in nine countries (Guinea, Liberia, Sierra Leone, Mali, Nigeria, Senegal, Spain, United Kingdom and the United States of America), with a total of 10,179 deaths in the first three West African countries experiencing widespread and intense transmission (http://apps.who.int/ebola/current-situation/ebola-situation-report-18-march-2015).

1 Materials and methods

Full-length EBOV genomes identified from the 2014 West African outbreak (n=81; 78 from Sierra Leone and three from Guinea) and previously identified EBOV genomes (n=26) were downloaded from GenBank and aligned using ClustalOmega [2] (Table S1 in Supporting Information). The alignment was 18,960 bp in length.

To gain insight into whether the nucleotide substitutions in the non-coding regions bear phylogenetic and evolutionary information, we compiled three different datasets, the full-length genome sequences of EBOV, concatenated coding regions, and the glycoprotein (GP) sequences. The GP protein of EBOV is responsible for receptor binding and fusion of the viral and cellular membranes. Recently, it has been reported to be able to mediate the level of certain microRNAs [3], which could be served as potential targets for Ebola drugs. Phylogenic analyses were performed using a Bayesian evolutionary analysis by sampling trees (Beast) approach [4], and 24 different combinations of molecular clock models (n=3) and coalescent models (n=8) were applied (Table S2 in Supporting Information). Fifty million steps were run and the first five million steps were removed as burn-in. Results obtained from different model combinations were compared using Bayes factor estimator of the marginal likelihood [5]. Maximum clade credibility trees were generated using TreeAnnotator v1.8 with a burn-in

*Corresponding author (email: wfshi@tsmc.edu.cn)

rate of 10%. Statistical analyses were performed using the Wilcoxon rank-sum test.

2 Results and discussion

Previous studies have revealed that EBOV formed a "ladder-like" phylogenetic structure and viruses collected from different time points clustered together comprising each "ladder"/clade of the tree [6]. In addition, the 2014 EBOV variants from West Africa formed a novel discrete clade in the phylogenetic tree [7]. Based on collection times, we classified the 107 sequences into five groups: 1976, 1995, 2002, 2007 and 2014. Nucleotide substitution events were compared with the adjacent groups (Table 1) and we found that nucleotide substitutions occurring in the non-coding regions accounted for 34.8% of all nucleotide mutations occurring between groups 1976 and 1995. The percentages of nucleotide substitutions occurring in the non-coding regions between other groups were even higher (Table 1). In particular, from group 1996 to group 2002, approximately 41% (172 out of 422) of the nucleotide mutations occurred in the non-coding regions (Table 1).

Bayes factor tests failed to support the results calculated using the strict clock model (Tables S3–S5 in Supporting Information). Therefore, only results obtained using the exponential relaxed clock and lognormal relaxed clock were further analyzed. Our phylogenetic analysis also revealed a "ladder-like" phylogenetic structure (Figure 1A), consistent with previous studies [6,7]. Sequences in the discrete lineages shown in Figure 1A corresponded well to those of the groupings in the first step. However, with regards to the higher resolution phylogenetic classification of the EBOV identified from Sierra Leone and Guinea in the ongoing 2014 outbreaks, our phylogenetic analysis revealed four distinct phylogenetic topologies (Figure 1B–E). In Figure 1B, the three EBOVs from Guinea formed a cluster paraphyletic to the cluster consisting of 78 EBOV identified in Sierra Leone in 2014, which suggested that introduction of the viruses into Guinea and Sierra Leone might be not related. In Figure 1C–E, although the three phylogenetic structures were slightly different in placing the three viruses from Guinea, all of them supported that the EBOV responsible for the Sierra Leone outbreak stemmed from those of Guinea. However, based on current surveillance data, the 2014 EBOV outbreak first emerged in Guinea in March, 2014 [1,7]. The first 12 patients with EBOV from Sierra Leone were believed to have attended the funeral of an Ebola virus disease case from Guinea [7]. Therefore, the phylogeny displayed in Figure 1B was not concordant with surveillance data, whereas phylogenies revealed in Figure 1C–E were generally consistent with surveillance data.

We next compared the phylogenies obtained using the exponential relaxed clock and lognormal relaxed clock (Table 2). As can be seen from Table 2, phylogenies obtained from 12 out of 16 phylogenetic trees calculated using the full-length EBOV genome sequences were in agreement with surveillance data. However, none of the 16 phylogenetic trees estimated using only the coding regions was able to reproduce the phylogenies and viral transmission data obtained from whole-genome and epidemiological analyses during the 2014 outbreak. With regards to the analysis based on the GP encoding region, six out of 16 phylogenetic trees supported the Sierra Leone outbreak stemmed from the Guinea outbreak. Therefore, it is apparent that only where the full-length EBOV genome sequences were employed that phylogenies could be obtained to reveal the actual evolutionary history and transmission route of the 2014 EBOV

Table 1 Distribution of the nucleotide substitution events across the full-length EBOV genome from 1976 to 2014

	Genetic region	Start point	End point	Group 1976→ Group 1996	Group 1996→ Group 2002	Group 2002→ Group 2007	Group 2007→ Group 2014
Inter-genic regions	3′ terminus→NP	1	469	3	12	16	23
	NP→VP35	2,690	3,128	3	15	8	12
	VP35→VP40	4,152	4,478	3	13	11	9
	VP40→GP	5,460	6,038	11	36	31	32
	GP→VP30	8,069	8,508	4	17	8	22
	VP30→VP24	9,376	10,344	14	38	38	44
	VP24→L	11,101	11,580	5	15	11	16
	L→5′ terminus	18,220	18,959	11	26	20	40
Sum				54	172	143	198
Coding regions	NP	470	2,689	27	35	41	55
	VP35	3,129	4,151	4	15	21	16
	VP40	4,479	5,459	7	15	13	15
	GP	6,039	8,068	20	45	45	54
	VP30	8,509	9,375	2	13	13	9
	VP24	10,345	11,100	3	13	12	18
	L	11,581	18,219	38	114	113	166
Sum				101	250	258	333

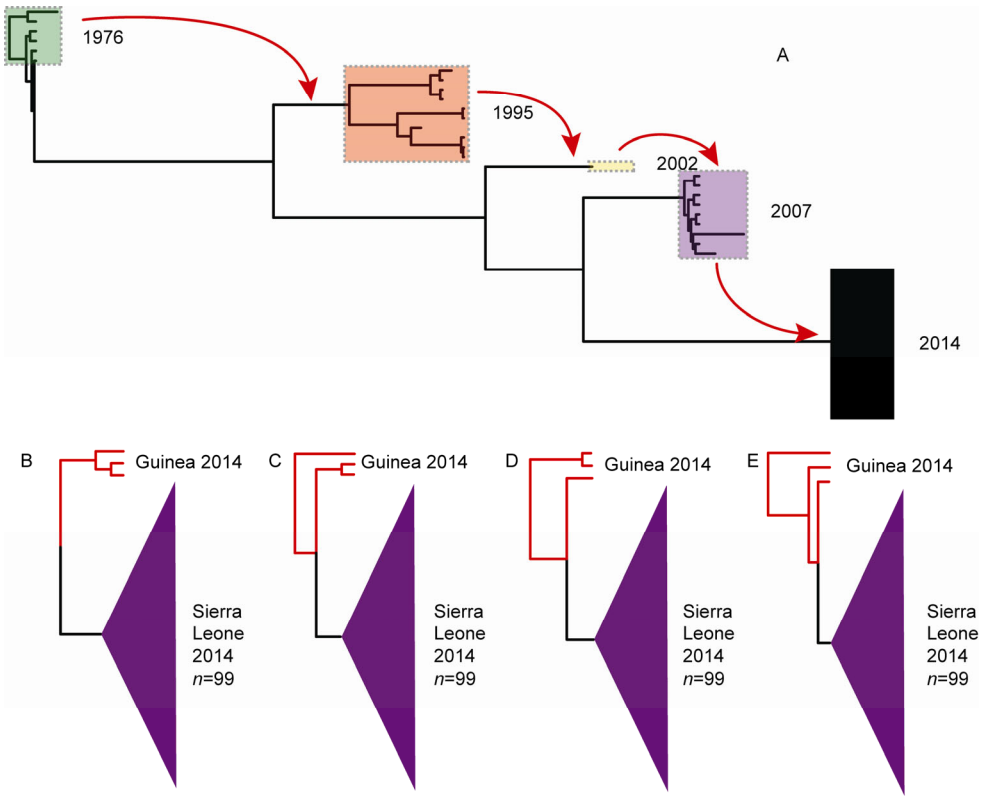

Figure 1 Panel A represents a schematic phylogenetic structure of EBOV. The black box of 2014 means the phylogenetic classification of the EBOV identified in the 2014 outbreak remains uncertain. Panels B–E display the four different phylogenetic topologies regarding the phylogenetic classification of the EBOV identified in the 2014 outbreak.

Table 2 Phylogenetic structures for EBOV genomes obtained from different model combinations and datasets

Model combinations		Datasets		
Molecular clock model	Coalescent model	Full-length genome sequences	Concatenated coding regions	GP
Uncorrelated exponential relaxed clock	Constant size	Figure 1B	Figure 1B	Figure 1B
	Exponential growth	Figure 1C	Figure 1B	Figure 1B
	Logistic growth	Figure 1C	Figure 1B	Figure 1B
	Expansion growth	Figure 1C	Figure 1B	Figure 1D
	Bayesian skyline	Figure 1C	Figure 1B	Figure 1E
	Extended Bayesian skyline plot	Figure 1C	Figure 1B	Figure 1E
	GMRF Bayesian skyride	Figure 1B	Figure 1B	Figure 1B
	Bayesian SkyGrid	Figure 1B	Figure 1B	Figure 1B
Uncorrelated lognormal relaxed clock	Constant size	Figure 1C	Figure 1B	Figure 1B
	Exponential growth	Figure 1C	Figure 1B	Figure 1B
	Logistic growth	Figure 1C	Figure 1B	Figure 1B
	Expansion growth	Figure 1C	Figure 1B	Figure 1B
	Bayesian skyline	Figure 1D	Figure 1B	Figure 1D
	Extended Bayesian skyline plot	Figure 1C	Figure 1B	Figure 1E
	GMRF Bayesian skyride	Figure 1B	Figure 1B	Figure 1B
	Bayesian SkyGrid	Figure 1C	Figure 1B	Figure 1D

outbreak.

It should be noted that the topology in Figure 1E has been reported in a previous study; however, the posterior probability to support this phylogeny was only 0.73 [7]. In our analyses, only using model combinations to calculate the GP dataset could we obtain this scenario. However, when the full length genome sequences were used, 11 out of 16 phylogenetic trees were Figure 1C-like, some of which gained higher support from Bayes factor test. Therefore, uncertainty still existed to some degree regarding the precise phylogenetic classification of the three Guinean variants and the transmission of the novel EBOV from Guinea

to Sierra Leone.

We then compared three evolutionary parameters estimated from different model combinations from different datasets: mean rate of EBOV, Kappa, and time to the most recent common ancestor (tMRCA) for the 2014 EBOV (Table S6 in Supporting Information). Mean rate (also named as μ) was used to compare the evolutionary rate [8,9] and Kappa (also named as κ) denotes the transition/transversion ratio. Both of them are important evolutionary parameters in several bioinformatics and evolutionary programs [10,11]. The estimate of tMRCA is often used to date the origin and transmission of a viral pathogen [7,12,13]. As values of the three parameters calculated using the Gaussian Markov random field (GMRF) Bayesian skyride model clearly deviated from the reasonable interval, they were removed from subsequent statistical analyses.

Statistical analyses showed that the three parameters estimated from the full-length genome sequences were significantly different from those from the coding regions (Figure 2). For the values calculated from the full-length genome sequences and GP sequences, they were significantly different in the kappa values, but not significantly different in mean rate and tMRCA (Figure 2). Therefore, the non-coding regions of the EBOV genome include indispensable evolutionary information and play an important role in estimating various evolutionary parameters. Neither the complete coding regions nor the GP region alone could reproduce accurate estimates of evolutionary parameters derived from whole genome analysis.

In summary, non-coding regions of EBOV genome contain indispensable phylogenetic and evolutionary information, and omitting nucleotide substitution information present in these regions can give rise to unreliable and even erroneous phylogenies and an inaccurate estimate of evolutionary parameters. Therefore, full-length EBOV genome sequences are recommended for the purpose of phylogenetic analysis and calculation of evolutionary parameters.

Full length genome vs Coding regions

	Mean rate	Kappa	tMRCA
Z	–2.412	–4.824	–2.525
P	0.015	0	0.011

Full length genome vs GP

	Mean rate	Kappa	tMRCA
Z	–0.302	–4.824	–0.641
P	0.780	0	0.539

Figure 2 Wilcoxon rank-sum test for EBOV full-length genome versus sub-genomic fragment analysis of the three evolutionary parameters: mean rate, Kappa and tMRCA.

This work was supported by the National Natural Science Foundation of China (81470096) and the Doctoral Starting up Foundation of Taishan Medical College. Shi WeiFeng was also supported by a grant from the International Development Research Centre.

1 Baize S, Pannetier D, Oestereich L, Rieger T, Koivogui L, Magassouba N, Soropogui B, Sow MS, Keïta S, De Clerck H, Tiffany A, Dominguez G, Loua M, Traoré A, Kolié M, Malano ER, Heleze E, Bocquin A, Mély S, Raoul H, Caro V, Cadar D, Gabriel M, Pahlmann M, Tappe D, Schmidt-Chanasit J, Impouma B, Diallo AK, Formenty P, Van Herp M, Günther S. Emergence of Zaire Ebola virus disease in Guinea—preliminary report. N Engl J Med, 2014, 371: 1418–1425

2 Sievers F, Wilm A, Dineen DG, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol, 2011, 7: 539

3 Sheng M, Zhong Y, Chen Y, Du J, Ju X, Zhao C, Zhang G, Zhang L, Liu K, Yang N, Xie P, Li D, Zhang MQ, Jiang C. Hsa-miR-1246, hsa-miR-320a and hsa-miR-196b-5p inhibitors can reduce the cytotoxicity of Ebola virus glycoprotein *in vitro*. Sci China Life Sci, 2014, 7: 959–972

4 Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol, 2012, 29: 1969–1973

5 Suchard MA, Weiss RE, Sinsheimer JS. Bayesian selection of continuous-time Markov chain evolutionary models. Mol Biol Evol, 2001, 18: 1001–1013

6 Walsh PD, Biek R, Real LA. Wave-like spread of Ebola Zaire. PLoS Biol, 2005, 3: e371

7 Gire SK, Goba A, Andersen KG, Sealfon RS, Park DJ, Kanneh L, Jalloh S, Momoh M, Fullah M, Dudas G, Wohl S, Moses LM, Yozwiak NL, Winnicki S, Matranga CB, Malboeuf CM, Qu J, Gladden AD, Schaffner SF, Yang X, Jiang PP, Nekoui M, Colubri A, Coomber MR, Fonnie M, Moigboi A, Gbakie M, Kamara FK, Tucker V, Konuwa E, Saffa S, Sellu J, Jalloh AA, Kovoma A, Koninga J, Mustapha I, Kargbo K, Foday M, Yillah M, Kanneh F, Robert W, Massally JL, Chapman SB, Bochicchio J, Murphy C, Nusbaum C, Young S, Birren BW, Grant DS, Scheiffelin JS, Lander ES, Happi C, Gevao SM, Gnirke A, Rambaut A, Garry RF, Khan SH, Sabeti PC. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science, 2014, 345: 1369–1372

8 Li YH, Chen SP. Evolutionary history of Ebola virus. Epidemiol Infect, 2014, 142: 1138–1145

9 Dudas G, Rambaut A. Phylogenetic analysis of Guinea 2014 EBOV Ebolavirus outbreak. PLoS Curr, 2014, doi: 10.1371/ currents. outbreaks.84eefe5ce43ec9dc0bf0670f7b8b417d

10 Bedford T, Cobey S, Beerli P, Pascual M. Global migration dynamics underlie evolution and persistence of human influenza A (H3N2). PLoS Pathog, 2010, 6: e1000918

11 Li XB, Zhang Z, Yu AL, Ho SYW, Carr MJ, Zheng WM, Zhang Y, Zhu C, Lei F, Shi W. Global and local persistence of influenza A(H5N1) virus. Emerg Infect Dis, 2014, 20: 1287–1295

12 Shi WF, Shi Y, Wu Y, Liu D, Gao GF. Origin and molecular characterization of the human-infecting H6N1 influenza virus in Taiwan. Protein Cell, 2013, 4: 846–853

13 Liu D, Shi WF, Shi Y, Wang DY, Xiao HX, Li W, Bi Y, Wu Y, Li X, Yan J, Liu W, Zhao G, Yang W, Wang Y, Ma J, Shu Y, Lei F, Gao GF. Origin and diversity of novel avian influenza A H7N9 viruses causing human infection: phylogenetic, structural, and coalescent analyses. Lancet, 2013, 381: 1926–1932

Supporting Information

Table S1 Full length EBOV genome sequences used in this study

Table S2 The 24 different model combinations used in this study

Table S3 Bayes factor test of the results obtained using the 24 model combinations for the full length genome sequences

Table S4 Bayes factor test of the results obtained using the 24 model combinations for the coding regions

Table S5 Bayes factor test of the results obtained using the 24 model combinations for the GP sequences

Table S6 Evolutionary parameters estimated using different model combinations

The supporting information is available online at life.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.